Volume 13 • Number 4 • 2010
VALUE IN HEALTH

Analysis of Health Utility Data When Some Subjects Attain the Upper Bound of 1: Are Tobit and CLAD Models Appropriate?

Eleanor M. Pullenayegum, PhD,[1,2] Jean-Eric Tarride, PhD,[1,3] Feng Xie, PhD,[1,3] Ron Goeree, MA,[1,3] Hertzel C. Gerstein, MD,[1] Daria O'Reilly, PhD[1,3]

[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada; [2]Biostatistics Unit, St Joseph's Healthcare Hamilton, Hamilton, ON, Canada; [3]Programs for Assessment of Technology in Health, St Joseph's Healthcare Hamilton, Hamilton, ON, Canada

ABSTRACT

Background: Health utility data often show an apparent truncation effect, where a proportion of individuals achieve the upper bound of 1. The Tobit model and censored least absolute deviations (CLAD) have both been used as analytic solutions to this apparent truncation effect. These models assume that the observed utilities are censored at 1, and hence that the true utility can be greater than 1.We aimed to examine whether the Tobit and CLAD models yielded acceptable results when this censoring assumption was not appropriate.

Methods: Using health utility (captured through EQ5D) data from a diabetes study, we conducted a simulation to compare the performance of the Tobit, CLAD, ordinary least squares (OLS), two-part and latent class estimators in terms of their bias and estimated confidence intervals. We also illustrate the performance of semiparametric and nonparametric bootstrap methods.

Results: When the true utility was conceptually bounded above at 1, the Tobit and CLAD estimators were both biased. The OLS estimator was asymptotically unbiased and, while the model-based and semiparametric bootstrap confidence intervals were too narrow, confidence intervals based on the robust standard errors or the nonparametric bootstrap were acceptable for sample sizes of 100 and larger. Two-part and latent class models also yielded unbiased estimates.

Conclusions: When the intention of the analysis is to inform an economic evaluation, and the utilities should be bounded above at 1, CLAD, and Tobit methods were biased. OLS coupled with robust standard errors or the nonparametric bootstrap is recommended as a simple and valid approach.

Keywords: CLAD, economic evaluation, nonnormality, Tobit, Utility.

1. Introduction

The quality-adjusted life-year (QALY) is a preferred outcome measure in economic evaluations as it incorporates both quantity and quality of life, thus allowing for comparisons across diseases [1,2]. The QALY can be calculated by multiplying the quality weight of a health state by the time spent in that state. Utility can be used as quality weight in economic evaluations and it is commonly estimated using existing multi-attribute health classification systems (also known as preference-based health-related quality of life (HRQoL) instruments), for example the Health Utility Index (HUI) [3,4] and the EQ5D [5]. In addition to preference-based instruments, there are profile-based instruments (either generic or disease-specific) used to measure HRQoL.

When analyzing HRQoL or a health utility, it is important to check that the distributional assumptions of the analysis method are met by the data. A common property of utility data is that the distribution is often nonnormal, with a left-skew and apparent truncation at 1 (i.e., a proportion of the population may achieve the upper bound of 1, depending on the population of interest). Simple linear regression is a common analysis technique which requires that the residuals of the regression be normally distributed. In the case of substantial truncation or bimodality of the utilities themselves, normality of the residuals may not hold. Alternative analytic techniques are Tobit models [6,7], censored least absolute deviations (CLAD) models [6,7], two-part models (TPMs), and latent class models (LCMs) for the analysis of such data [8]. Simulation studies have compared these methods head-to-head (ordinary least squares (OLS), Tobit, and CLAD in [6], OLS, CLAD, two-part, and LCMs in [8]). However, the CLAD and Tobit approaches model something quite different to the other models, and previous comparisons between the methods have often not discussed the key conceptual differences between them (see, for example [8]).

Whereas the two-part and LCMs take various approaches to dealing with the conditional nonnormality in the data, the Tobit and CLAD models treat the distribution as censored at 1. That is, they assume that the underlying measurement can extend beyond 1, but that the measurement instrument incorrectly truncates the distribution at 1. In this article, we shall make a clear distinction between health utilities and HRQoL, and will argue that health utilities should not usually be treated as censored. We will show through a simulation study that linear regression coupled with the nonparametric bootstrap is quite adequate for the analysis of health utility data. We will also compare the performance of the OLS, Tobit, CLAD, two-part, and LCMs.

The remainder of this article is organized as follows: Section 2 outlines the philosophical differences between a health utility and HRQoL and discusses some measurement issues; Section 3 describes some analysis approaches; Section 4 describes the set-up of a simulation study to illustrate the importance of selecting a method that is appropriate to the desired outcome; Section 5 presents the results of the study; and Section 6 offers some conclusions.

2. Concepts and Measurement

For the purposes of this article, it is helpful to make a clear distinction between a health utility and HRQoL. Before discussing

Address correspondence to: Eleanor M. Pullenayegum, St Joseph's Healthcare Hamilton, 50 Charlton Ave. E, Hamilton, ON L8N 4A6, Canada. E-mail: pullena@mcmaster.ca

10.1111/j.1524-4733.2010.00695.x

analysis issues, we shall first discuss whether it is a health utility or HRQoL that is the outcome of interest, and then consider measurement issues.

Health utility is a quality weight used to calculate a QALY. Health utilities can be derived using different elicitation techniques, including the visual analog scale, standard gamble (SG), and time trade-off (TTO), with 0 representing death and 1 representing full health. These elicitation techniques ensure that the weights have ratio properties, i.e., a utility of 0.5 is worth half as much as a utility of 1. Negative utilities are allowed because there are health states worse than death; however, utilities cannot exceed 1 because you cannot do better than full health [9].

In contrast, HRQoL is a broader and a more abstract construct that captures an individual's well-being. There is no bound set to HRQoL measurements. Different instruments developed over different scales will result in different score boundaries. For example, the Short Form 36 (SF-36) bodily pain domain score ranges from 0 to 100 [10], while the Western Ontario and McMaster Universities Osteoarthritis Index pain domain score ranges from 0 to 20 [11].

Turning now to the measurement issue, many instruments used to capture health utilities have the property that a portion of the population achieves the upper bound of 1. This phenomenon is often described in the literature as "censoring," "truncation," or a "ceiling effect." Implicit in each of these terms is that the true measurement can exceed 1. However, this is not the only manner in which the data distribution can arise. An alternate explanation is that the measurement instrument is not able to capture small departures from perfect health.

It is not possible to capture all health states using a short questionnaire. For example, the EQ5D asks just five questions. While it is clearly impossible to assess a person's well-being comprehensively using a generic instrument and based on such little information, accurate measurement must be balanced with the need to make questionnaires simple enough to be practical. If we are interested in a utility, we must recognize that these measures are approximations only and are insensitive to small departures from perfect health. We stress, however, that this is a measurement issue rather than a censoring issue. In healthy populations, instead of seeing a large number of values clustered slightly below one, we see a large number of values that are exactly one. If values were censored at one, we would see a proportion of values that achieve that bound, but it would be possible to exceed that bound. For a utility, this is by definition impossible: you cannot exceed full health. On the other hand, if we are interested in HRQoL as a more abstract construct, then we can acknowledge that our measurement scale does not distinguish well between normal and perfect health, and hence we can take values of one to indicate normal health and treat the measure as censored, so that some individuals can have "supranormal" health [12].

Analysis decisions become clearer once it has been decided whether it is a utility or the HRQoL construct that is of interest. Although HRQoL is often an important outcome in its own right, in the context of an economic evaluation in which costs are to be weighed against effects and the measure of effectiveness is a QALY (i.e., a cost–utility analysis in particular), it is the utility rather than HRQoL that is the outcome of interest. We now describe appropriate methods of analysis for each outcome.

3. Analysis Methods

Here, we review the various methods that have been proposed for the analysis of health utility data. There are two broad categories of models: the first category assumes the data is censored at 1, while the second assumes no censoring but must address the nonnormality of the data. The methods are outlined in Table 1, and summarized below.

The Tobit and CLAD models assume that the true values can extend beyond one but are observed subject to censoring at one. The Tobit model handles this censoring by assuming that the true value has a normal distribution whose mean is given by a linear combination of the covariates. The CLAD model assumes that the median is a linear combination of the covariates, but leaves the distribution otherwise unspecified. When making the censoring assumption, it is implicit that it is HRQoL rather than a health utility that is the outcome of interest.

When it is a utility that is the outcome of interest, the assumption is that the true utility score has a maximum of 1. Thus the question is not how to deal with the censoring, but rather how to deal with the nonnormality. OLS applied to marginal linear regression models handles this by modeling the mean utility only, and makes no assumptions about the remainder of the distribution. TPMs specifically model the probability of attaining the upper bound, and then model the remainder of the distribution below this bound using a regression model, sometimes after applying a log transform. LCMs assume that there is an unobserved variable which splits the population into two, and that the distribution of utilities within each population is normal. TPMs may thus be more helpful when a large proportion of the population achieve a utility of 1, whereas LCMs may be more useful when the distribution of observed utilities is bimodal, with both modes falling below 1.

We now describe each method in detail.

Models Assuming Censoring, and Hence Candidates for the Analysis of HRQoL Subject to Ceiling Effects

Tobit. The first model that we shall consider is the Tobit model. This was developed by Tobin, and assumes that there is a latent HRQoL Y_i^* satisfying

$$Y_i^* = X_i\beta + \varepsilon_i^*$$

with $\varepsilon_i^* \sim N(0, \sigma^2)$, but that instead of observing Y_i^*, we instead observe Y_i^*, defined as

$$Y_i = Y_i^* \text{ if } Y_i \leq 1$$

$$Y_i = 1 \text{ otherwise}$$

This is the same model as an accelerated failure time model and can thus be fitted by any procedure that fits a parametric survival analysis using a normal distribution for the failure time. In the Tobit model, it is the latent HRQoL Y_i^*, rather than the utility Y_i that is modeled.

In a simulation study, Austin et al. [7] demonstrated that the Tobit model is less biased than OLS when the error terms are homoscedastic (even if they are nonnormal). We note, however, that this was because the data were simulated so that the true health status could be larger than 1, depended linearly on the covariate (age), but was censored at 1. Thus, OLS was biased when the "true" regression coefficient was that used for the simulation, i.e., for health states that could be supranormal. Whether or not OLS is actually biased depends on whether you consider health states larger than 1 to be valid and of interest. Austin et al. [7] caution that the Tobit model can give misleading results in the presence of heteroscedascity.

Censored least absolute deviations. The CLAD approach is a solution to the Tobit model's sensitivity to heteroscedascity, and

Table I Models for health utility and health-related quality of life data

Model	Ceiling effects	Assumptions	Useful for	Statistical software
Tobit	True utility can extend beyond 1, but has been censored at 1	Underlying HRQoL is a linear function of the covariates plus a residual; residuals are Gaussian and homoscedasctic	Analyzing underlying HRQoL	SAS: proc qlim R/Splus: survreg STATA: Tobit
CLAD	True utility can extend beyond 1, but has been censored at 1	Underlying HRQoL has a median that is a linear function of the covariates	Analyzing HRQoL particularly when nonnormality or heteroscedasscity is suspected	R/Splus: crq in library quantreg STATA: CLAD
Marginal linear model	Does not need to be handled, because model is for the mean only	Marginal mean is a linear function of the covariates	Analyzing utilities, particularly when a linear model for the marginal mean is desired	SAS: proc reg R/Splus: lm and boot STATA: regress
Two-part	Point mass at 1 models the probability of attaining the upper bound	Log odds of attaining a utility of 1 is a linear function of covariates; for those utilities less than 1 the decrement from 1 is log-normal with a mean that is a linear function of the covariates	Analyzing utilities, especially when a substantial proportion of the population achieve the upper bound of 1.	SAS: proc genmod and proc reg R/Splus: glm and lm STATA: logit and regress
Latent class	There are two latent classes so that the resulting distribution is bimodal	Within each latent class, distribution of utilities is a linear function of the covariates plus a Gaussian, homoscedasctic residual	Analyzing utilities, especially when fewer individuals achieve the upper bound of 1, but distribution of utilities remains bimodal	SAS: proc lca R/Splus: flexmix STATA: gllamm

CLAD, censored least absolute deviations; HRQoL, health-related quality of life.

has been shown to yield consistent estimates under certain conditions. Like the Tobit model, CLAD assumes that HRQoL values measured to be 1 have in fact been censored, and accounts for this in the estimation. In contrast to the Tobit model, CLAD models medians rather than means: rather than minimizing a sum of squares, it minimizes the sum of absolute deviations given by

$$\sum_{i=1}^{n} |Y_i - \min(1, X_i\beta)|$$

Note that if interest lies in modeling QALYs in order to conduct an economic evaluation, it is the mean, not the median, that is of interest [13]. However, if the error distribution of the uncensored utility values is symmetric, the mean and the median will coincide.

Previous work [6] has recommended the use of CLAD over OLS, as it has a smaller absolute prediction error. However, because CLAD is a median regression and OLS is a mean regression, it is not surprising that CLAD should have a smaller absolute prediction error—the question is whether or not it has a smaller mean square prediction error. Work by Huang et al. [8] suggests that CLAD predicts a smaller percentage of both absolute and square error than OLS.

It is important to note that both the Tobit and the CLAD methods model the latent Y_i^*, i.e., a HRQoL, of which the observed utility is assumed to be an imperfect observation (specifically, an observation that is subject to censoring at 1).

Models with No Censoring, and Hence Candidates for the Analysis of Health Utilities

Linear regression, marginal mean models, and OLS. Conventional linear regression is fully parametric, that is, it places a distribution on the utilities. If Y_i is the measure of utility for individual i and X_i is a vector of covariates, then a typical linear regression model is

$$Y_i = X_i\beta + \varepsilon_i \tag{3.1}$$

with $\varepsilon_i^* \sim N(0, \sigma^2)$. This model thus assumes that the errors are normally distributed and homoscedasctic (i.e., that their variance is the same for all i). Under this standard linear regression model, Y, given X, is normal with mean $X\beta$ and variance σ^2. The regression coefficients β can be estimated by maximum likelihood, and so minimize

$$\sum_{i=1}^{n} (Y_i - X_i\beta)^2 \tag{3.2}$$

Hence $\hat{\beta}$ solves $\sum_{i=1}^{n} (Y_i - X_i\beta)X_i = 0$.

Normality of the distribution of $\hat{\beta}$ follows from conditional normality of Y, and this allows for calculation of P-values and confidence intervals. It is a familiar result that $\hat{\beta}$ is the best linear unbiased estimator (BLUE) for β.

However, the measured utility is bounded at 1, and hence the assumption that Y is normally distributed conditionally on covariates will not usually hold (this assumption can be checked by examining q–q plots of the residuals from the linear model). Moreover, unlike other outcomes (e.g., intelligence quotient, height) which are also bounded, because individuals will sometimes achieve the upper bound, conditional normality is often not even approximately achieved. A simple solution to this problem is to make a small change to the model. Rather than using the model (3.1), we instead take

$$E(Y_i|X_i) = X_i\beta \tag{3.3}$$

This is similar to (3.1) in terms of the marginal mean model; however, in this case we do not assume that the distribution of Y given X is normal. Motivated by the maximum likelihood estimator (3.2), we take $\hat{\beta}$ to minimize the sum of squares, i.e., as before $\hat{\beta}$ solves $\sum_{i=1}^{n}(Y_i - X_i\beta)X_i = 0$. This procedure is called OLS. Because model (3.3) makes fewer assumptions than model (3.1), $\hat{\beta}$ loses some of its properties. It is no longer guaranteed to be unbiased or normal, making calculation of *P*-values and confidence intervals more challenging. However, the central limit theorem can be used to show that $\hat{\beta}$ is both asymptotically unbiased (i.e., consistent) and asymptotically normal. Note that because the estimation procedures for linear regression and the marginal model are identical, any procedure for linear regression will fit the marginal model; however, the resulting *P*-values and confidence intervals will be valid only if the sample size is sufficiently large and the error terms are identically distributed.

Bootstrap. This last statement poses two difficulties. First, the term "sufficiently large" is not well defined. The answer to the question "How large is large enough?" depends on to what degree the assumptions of the conventional linear regression model are violated. At one extreme, if the data is perfectly normal, the only sample size requirement is that there are sufficiently many observations to fit the required model, and for the sample to be representative of the population of interest. At the other extreme, an outcome variable that is heavily skewed may require a sample size of thousands before the sampling distribution of $\hat{\beta}$ begins to approach normality. Generally speaking, the greater the degree of nonnormality, the larger the required sample size for asymptotic results to begin to apply.

Second, error terms from a linear regression of utility data are unlikely to be identically distributed, because individuals with higher scores also tend to have scores that are less variable, so that the data shows heteroscedascity. This is a consequence of the ceiling effect: individuals with normal but imperfect health will tend to have utilities close to 1, whereas individuals with poorer health could take on a wider range of utilities less than 1.

A useful method of deriving valid *P*-values and confidence intervals in the face of these difficulties is the bootstrap. There are two approaches to bootstrapping results from a marginal mean regression: the nonparametric bootstrap and the semiparametric bootstrap. A detailed description of the bootstrap is outside the scope of the current article; however, a good overview may be found in a previous study [14]. Briefly, the nonparametric bootstrap quantifies the sampling distribution of the statistic of interest (e.g., a regression coefficient) by creating a large number of simulated datasets, each of which is used to estimate the statistic of interest. The simulated datasets are created by sampling (with replacement) individuals from the actual dataset. One drawback to the nonparametric bootstrap is that it does not fix the distribution of the covariates, and so researchers may sometimes consider the semiparametric bootstrap. In the semiparametric bootstrap, instead of sampling the actual responses and covariates from the original dataset, the bootstrap samples the residuals from the linear model and fixes the covariates. This has the advantage of conditioning on the covariates; however, its validity relies on being able to assume that the residuals from the linear model are independent and identically distributed. The key problem for utility data is that the residuals are unlikely to be identically distributed because of the heteroscedascity in the data. We illustrate this problem in the simulation study below. Although work by Walters and Campbell [15] suggests that the bootstrap yields similar results to *t*-tests and linear regressions, we note that this was based on large datasets using individual dimensions of the SF-36, which do not represent utilities.

There are a number of options for constructing confidence intervals based on bootstrap re-samples (see Kenward and Carpenter [14] for a review). Here we shall consider two options: using the bootstrap standard errors in place of the model-based standard errors, and, in the case of the nonparametric bootstrap, using the bias-corrected and accelerated (bca) bootstrap confidence intervals.

Robust standard errors. Robust standard error estimation poses an alternative to the bootstrap when calculating *P*-values and confidence intervals. This method is designed to deal with heteroscedascity in the data and is suitable when sample sizes are sufficient for the regression coefficients to be approximately normal. The variance of the estimated regression coefficients is given by

$$\text{var}(\hat{\beta}) = \left(\sum_i X_i X_i'\right)^{-1}\left(\sum_i X_i var(Y_i|X_i)X_i'\right)\left(\sum_i X_i X_i'\right)^{-1}$$

When heterosceascticity is present, $\text{var}(Y_i|X_i)$ is estimated by a function of the residual $Y_i - X_i\beta$. The resulting covariance matrix for $\hat{\beta}$ is termed a heteroscedascticity consistent covariance matrix (HCCM). There are a number of versions; however, here we shall use the HCCM3 [16], which is designed to avoid observations with large variances over-contributing to the analysis:

$$\text{var}(\hat{\beta}) = \left(\sum_i X_i X_i'\right)^{-1}\left(\sum_i X_i\left(\frac{y_i - X_i\beta}{1-h_{ii}}\right)^2 X_i'\right)\left(\sum_i X_i X_i'\right)^{-1}$$

$$\text{where } h_{ii} = X_i'\left(\sum_j X_j X_j'\right)^{-1} X_i$$

TPMs and LCMs. If the researcher does not want to consider supranormal health states, an alternative to marginal linear models is to model the distribution of observed utilities. Huang et al. [9] consider TPMs and LCMs. The idea of both classes of model is to divide the population into two parts, the first of which tend to have higher utilities, and the second have lower utilities, so that the distribution of utilities in the total population is a mixture distribution. In the TPM, the first segment of the population has utilities equal to 1, whereas in the LCM, both segments have nontrivial distributions, and are distinguished from one another by a latent classification variable.

TPMs. Specifically, the TPM assumes that if Y_i is the utility for individual i, then

$$\text{logit}(P(Y_i = 1|X_i)) = X_i\alpha$$

$$E(Y_i|X_i, Y_i < 1) = X_i\beta$$

The TPM with a log transform explicitly acknowledges that if an individual does not have $Y = 1$ then $Y < 1$ by using a log-transform, namely

$$\log(1 - Y_i)|(Y_i < 1, X_i) \sim N(X_i\beta, \sigma^2)$$

LCMs. LCMs assume that observations fall into two or more unobserved (latent) classes, and then model the distribution within each class. In the context of utility data, which is frequently bimodal, it will often be appropriate to assume that there are two latent classes, and to adopt a normal distribution within each class. If C_i is the latent class variable for individual i, with $C_i \in \{1,2\}$, the latent class model would be

$$P(C_i = 2) - p_i$$

$$Y_i \sim N(X_i\beta_1, \sigma_1^2) \text{ if } C_i = 1$$

$$Y_i \sim N(X_i\beta_2, \sigma_2^2) \text{ if } C_i = 2$$

Estimation of these models is typically via the expectation–maximization algorithm (EM) [17].

Work by Huang et al. [8] suggests that for the specific dataset studied, the TPM (with a log transform) and the LCM perform the better than OLS and the TPM without a log transform, which in turn perform better than CLAD, where the measures of performance are the proportion of absolute deviation and square deviation in the validation dataset explained by model predictions derived from the training dataset. It should be noted that the TPMs and LCMs are parametric approaches and each assume normality. For the untransformed TPM, it is assumed that the portion of the distribution below 1 follows a normal distribution. For the transformed TPM, it is assumed that the log-transformed portion of the distribution below 1 follows a normal distribution. For the LCM, it is assumed that the distributions within the classes are normally distributed. While for the untransformed TPM and LCM, lack of normality is unlikely to cause bias in the estimated regression coefficients, serious bias could result for the transformed TPM if inferences are to be made about mean utilities on the untransformed scale.

4. Simulation Methods

This simulation study will explore the performance of the OLS, Tobit, CLAD, two-part and latent class methods.

Dataset Used for Simulations

Diabetes Hamilton (Hamilton, ON, Canada) is a community-oriented program that provides information and resources to people with diabetes. At the time of registration, participants consent to having some of their self-reported data included in a registry and complete a five-page questionnaire, which is entered centrally. Approximately 1000 Diabetes Hamilton participants also completed the EQ5D at the time of registration. The following simulation used EQ5D data and clinical data supplied by 1143 registrants in Diabetes Hamilton. The US scoring algorithm was used to calculate the EQ5D utilities [18].

Simulation Model

For the purposes of illustration, we chose one to look at the relationship between EQ5D and insulin. This variable was chosen for two reasons: firstl, it is binary, making interpretation straightforward, and second, a reasonable proportion of patients reported using insulin, which made simulation more straightforward. The distribution of EQ5D in the Diabetes Hamilton dataset is shown in Figure 1. This distribution is bimodal and shows the usual ceiling effect, with 17% (193/1143) of the EQ5D achieving the upper bound of 1. The minimum observed value of the EQ5D is –0.03836. We used this dataset to simulate EQ5D measurements following the empirical distribution observed in the study. Specifically, we randomly sampled 50, 100, 200, and 500 patients from the database, and recorded their values of the EQ5D and insulin. We also used the entire sample of 1141 with valid EQ5D and covariate measurements (two individuals were missing covariate data and were thus excluded). In each of the 1000 simulations, we first selected 50 individuals from the dataset of size 50 by sampling with replacement, then selected

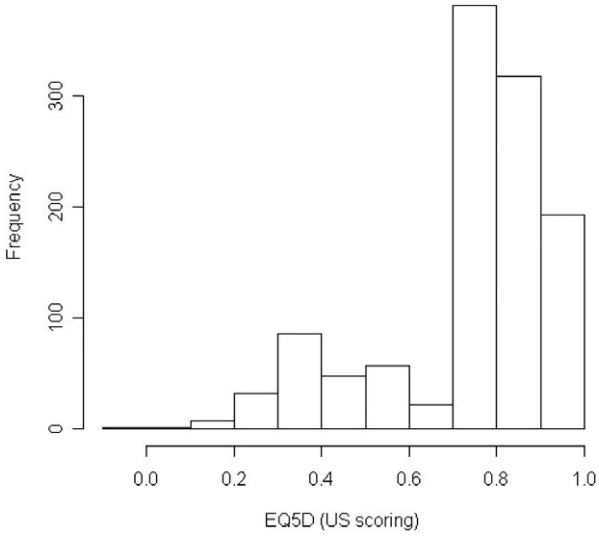

Figure 1 Histogram of EQ5D scores in the Diabetes Hamilton dataset, using US scoring.

100 individuals from the dataset of size 100, sampling with replacement, and similarly for the sample sizes of 200, 500, and 1141. This procedure amounts to sampling from the empirical distribution of the data for each given sample size, and is appealing as it avoids the need to make any distributional assumptions. We added a very small random error to the EQ5D measurements in each sampled datapoint (error ~ N(0, SD = 0.000001)) as this led to fewer singularities in fitting the CLAD model. When adding this random error led to EQ5D values that were greater than 1, these were truncated at 1.

We then computed the OLS, Tobit, CLAD, transformed two-part, untransformed two-part and latent class estimators of the unadjusted effect of taking insulin on utility, i.e., the estimators of the regression coefficient of insulin from a regression model of EQ5D onto an intercept and insulin alone. This coefficient is the mean difference in EQ5D for those on insulin versus those not taking insulin. Using the unadjusted coefficient is useful for the purposes of illustration, as in this simple case it is possible to calculate what the true regression coefficient for insulin should be (i.e., the mean difference in utility between those on insulin vs. those not on insulin); this is not the case when adjusting for other covariates.

Criteria for Assessing Model Performance

Three criteria can be used to compare the models: bias, confidence interval coverage probability, and estimated standard errors. These are described briefly below. Because the TPMs and LCMs do not provide estimates of standard errors of confidence intervals, these models are assessed in terms of their bias only.

Bias. The estimated bias in the coefficients of insulin is defined as the mean value of the coefficient across the 1000 samples minus the true value.

The true coefficient of insulin is derived by calculating the difference in mean EQ5D among those on insulin versus those not on insulin. The correct unadjusted coefficients of insulin are –0.008310555, –0.101493701, –0.089779487, and –0.066055266 for the subsamples of 50, 100, 200, and 500 patients, respectively; and –0.078575287 for the sample as a

Table 2 Bias (and bias/ses) statistics for the OLS, Tobit, CLAD, two-part, and latent class estimators

Sample Size	OLS	Tobit	CLAD	TPM trans	TPM No trans	LCM
50	−0.00068	−0.00252	−0.00108	−7439.15	−0.00068	−0.00068
	(−0.34)	(−1.17)	(−0.72)	(−1.30)	(−0.34)	(−0.34)
100	0.00012	−0.00327	0.05721	−29.92	0.00012	0.00014
	(0.11)	(−2.95)	(64.10)	(−9.26)	(0.11)	(0.13)
200	−0.00170	−0.01034	0.03497	−711.44	−0.00170	−0.00167
	(−1.77)	(−10.23)	(34.14)	(−16.39)	(−1.77)	(−1.74)
500	−0.00030	−0.00482	0.03433	−136.00	−0.00030	−0.00031
	(−0.47)	(−7.09)	(54.93)	(−24.32)	(−0.47)	(−0.48)
1141	0.00049	−0.00587	0.04225	−153.66	0.00049	0.00049
	(1.11)	(−12.76)	(85.67)	(−43.21)	(1.11)	(1.12)

The first number in each cell is the estimated bias, i.e., the mean estimate of the regression coefficient of insulin across the 1000 simulations, minus its true value. The second number (in parentheses) is the estimated bias divided by its standard error. If the estimator is unbiased, bias/se should be between −1.96 and +1.96 with probability 0.95. The numbers are bias with ratio of bias to standard error in parenthesis.
CLAD, censored least absolute deviations; LCM, latent class model; OLS, ordinary least squares; TPM, two-part model.

whole. It may seem surprising that all these coefficients are negative, so that patients on insulin have lower utilities than those who are not. However, this is an observational study and the coefficients are unadjusted, so it is likely that patients who require insulin therapy are in worse health than those who do not.

In examining bias statistics, clearly we would like the bias to be zero. Because we are limited to a finite number of simulations, our results will be subject to sampling variation, and so in addition to reporting bias, we report the bias divided by its standard error (which is equal to the standard error of the mean of the regression coefficients). If an estimation technique is unbiased, we would expect the observed bias divided by its standard error to lie between –1.96 and +1.96 95% of the time.

Coverage probability. In addition to calculating bias, we will also look at how adequate the 95% confidence intervals for the estimated coefficients of insulin are. Coverage probability is defined as the proportion of times that the estimated 95% confidence interval contains the true value of the regression coefficient. Then, if the confidence intervals are correct, we expect the coverage probability to be 0.95. Again, because we are limited to a finite number of simulations, we will not see exactly 95% coverage, but with 1000 simulations we would expect the observed coverage probability to lie between 93.6 and 96.4% 95% of the time. Coverage probabilities are available only for the linear regression, Tobit, and CLAD models, as the other methods do not return standard errors or confidence intervals for the estimated mean difference in utility between those on insulin versus those not on insulin.

Empirical standard errors (ESEs) and average estimated standard errors (ASEs). In order to understand any departures of the coverage probabilities from the required 0.95, it is helpful to look at estimates of standard error. For each estimator, the ESE is the standard deviation of the 1000 estimates of the unadjusted regression coefficient for insulin. The ASE is the average of the model-based standard errors for each of the 1000 simulations. If the model-based standard errors are correct, the ASE should be close to the ESE. As with the coverage probabilities, ASEs will be available only for the linear regression, Tobit and CLAD models.

5. Simulation Results

Table 2 shows that if we are interested in the association between insulin and utility, the Tobit, CLAD, and transformed TPM are biased, but the OLS, untransformed two-part, and latent class estimators are unbiased. The extent of the bias does not appear to depend on sample size. Bias becomes easier to detect for larger sample sizes because as the sample size increases, the standard error of the estimated bias decreases, so that when important bias is present the estimated bias divided by its standard error increases with increasing sample size.

Table 3 shows that the OLS, Tobit, and CLAD methods yield confidence intervals that have coverage probabilities that are in general too low, dramatically so for the CLAD. In the case of the CLAD method, this can be explained by the bias. In the case of the Tobit, the undercoverage could be due to bias, to heteroscedascity, or to nonnormality. In the case of OLS, the undercoverage could be due to either heteroscedascity or to nonnormality. Undercoverage of the confidence intervals corresponds to confidence intervals that are too narrow, and hence *P*-values that are anticonservative. For example, for the OLS estimator with a sample size of 1141, the coverage probability of 0.927 means that the model-based *P*-value would be 0.05 when the true *P*-value was 0.073. Thus, the error in the confidence intervals and *P*-values is in fact very serious and would lead to seriously inflated type I errors.

Examining the empirical and estimated standard errors of the regression coefficients helps to explain the observed coverage probabilities. Table 4 shows that for the OLS and Tobit estimators, the empirical standard errors are all larger than the estimated standard errors. Thus, the model-based standard errors are too small, which has led to confidence intervals that are too narrow.

We now explore the role of robust standard errors and bootstrapping for the marginal linear model. Semiparametric bootstrapping has been suggested for OLS estimators; however, it is not recommended as it leads to coverage probabilities that are too low (see Table 5). This is because semiparametric bootstrap-

Table 3 Coverage probabilities of the 95% confidence intervals for the OLS, Tobit, and CLAD estimators

Sample Size	OLS	Tobit	CLAD
50	0.939	0.939	0.660
100	0.922	0.924	0.217
200	0.950	0.942	0.725
500	0.941	0.943	0.272
1141	0.927	0.921	0.029

If the confidence intervals are correct, the coverage probabilities should all be 0.95.
CLAD, censored least absolute deviations; OLS, ordinary least squares.

Table 4 ESE and ASE for the OLS, Tobit, and CLAD methods

Sample Size	OLS			Tobit			CLAD		
	ESE	ASE	ASE/ESE	ESE	ASE	ASE/ESE	ESE	ASE	ASE/ESE
50	0.064	0.063	0.993	0.068	0.067	0.981	0.048	0.060	1.269
100	0.034	0.031	0.898	0.035	0.032	0.903	0.028	0.015	0.530
200	0.030	0.030	0.976	0.032	0.032	1.001	0.032	0.040	1.241
500	0.020	0.020	0.975	0.021	0.021	0.996	0.020	0.015	0.783
1141	0.014	0.013	0.926	0.015	0.014	0.950	0.016	0.007	0.475

The ESE is the standard deviation of the estimated regression coefficients over the 1000 simulations. The ASE is the standard error estimate for the regression coefficients returned by the model, averaged over the 1000 simulations. If the model-based estimate of the standard error is correct, the ASE and the ESE should be close.
ASE, average estimated standard error; CLAD, censored least absolute deviations; ESE, empirical standard error; OLS, ordinary least squares.

ping assumes that the residuals are homoscedasctic, which is not the case in this example. By contrast, the robust standard errors and nonparametric bootstrap bca intervals give acceptable coverage probabilities for the OLS estimator for sample sizes of 100 or larger, while replacing the model-based standard errors with the nonparametric bootstrap standard errors gives acceptable coverage probabilities for sample sizes of 200 or larger.

When it is health utilities rather than HRQoL that are of interest, linear regression is not the only option. Table 6 explores the relative efficiency, in terms of estimated standard errors, of the linear regression, untransformed two-part and a latent class estimators. All three methods have comparable variability, which is not surprising as in this case we have just a single binary covariate, and it is well known that the OLS estimators are BLUE.

6. Discussion

Previous work has advocated the use of Tobit or CLAD models for the analysis of HRQoL data when the primary focus is on regression modeling, understanding the relationship between HRQoL and a covariate, or on explaining variability in HRQoL [6,7]. This approach has, however, also been adopted when analyzing health utility data for the purposes of economic evaluation. In this context, Tobit and CLAD models are not appropriate as they assume that utilities can extend beyond 1, when in fact for the EQ5D or HUI health states are scored according to empirically estimated utilities, treating 1 as the maximum score. Our simulations show that when the utility must be bounded above at 1, the Tobit and CLAD models both lead to bias.

When utilities are not treated as having been observed subject to censoring, the analysis must deal with the conditional non-normality and heteroscedascity of the data. OLS gives unbiased estimators of regression parameters regardless of distributional assumptions; however, the estimated standard errors will usually not be correct even for large sample sizes because they are calculated based on the assumption of homoscedascity. A simple solution is to use a robust standard error or the nonparametric bootstrap. Some authors [15] have suggested a semiparametric or parametric bootstrap based on re-sampling the residuals; however, this is not appropriate as it assumes that the residuals all have equal variance. This is borne out by our simulation study, which shows that for our data, the semiparametric bootstrap led to confidence intervals that were too narrow.

While OLS coupled with the bootstrap or robust standard errors provides asymptotically unbiased regression coefficients and valid confidence intervals, the estimates may not be as precise as they could be. The reason for this is that OLS fits a marginal model, that is, a model for the mean utility, not for the whole distribution. Generally, if a good model for the data distribution could be found, one would expect that it would yield more precise estimates of the relevant regression coefficients. Various alternatives have been proposed, among them are beta distributions, TPMs, and LCMs.

Beta models are an appealing strategy because they assume that utilities are bounded above at 1; however, they also assume a lower bound of 0, which may not be appropriate for all populations. For example, in our population, the minimum EQ5D was in fact negative. TPMs are useful in describing the ceiling effect. The drawback, however, is that if we are interested in means, they are not direct outputs of the model, especially when more than one covariate is considered in the linear regression. In our simulation example, the transformed TPM was biased because the assumption of normality and homoscedascticity of the log-transformed utilities was not met. The untransformed TPM was unbiased and showed similar precision to linear regression. This precision result is, however, sensitive to the distribution of the data and hence will not necessarily hold in general.

LCMs are an attractive alternative when the distribution of utility scores is bimodal. There is thus nothing in the model that stipulates that utility scores should be bounded; however dividing the population into subgroups can help to ensure that the model-based probability of an individual falling outside of the bounds is small. Huang et al. found that TPMs and LCMs outperformed OLS in terms of minimizing residual sums of squares [8]. In our simulation example, LCMs were no more precise than

Table 5 OLS coverage probabilities using the model-based, robust, and semiparametric and nonparametric bootstrap standard errors, and the nonparametric bootstrap bca confidence intervals

Sample size	Model-based	Robust	Semiparametric bootstrap	Nonparametric bootstrap standard errors	Nonparametric bootstrap bca intervals
50	0.939	0.934	0.928	0.930	0.909
100	0.922	0.950	0.925	0.934	0.964
200	0.950	0.945	0.952	0.948	0.950
500	0.941	0.953	0.939	0.946	0.945
1141	0.927	0.944	0.925	0.940	0.94

bca, bias-corrected and accelerated; OLS, ordinary least squares.

Table 6 Comparison of empirical standard errors among methods appropriate for health utility data

Sample size	Untransformed		
	OLS	TPM	LCM
50	0.064	0.064	0.064
100	0.034	0.034	0.034
200	0.030	0.030	0.030
500	0.020	0.020	0.020
1141	0.014	0.014	0.014

These numbers are the observed standard deviation of estimated difference in utility for those using insulin versus those not using insulin among the 1000 simulated datasets. LCM, latent class model; OLS, ordinary least squares; TPM, two-part model.

linear regression in terms of the sample standard error, however as with the TPM, it is important to recognize that the relative efficiency of latent class and linear regression models will be specific to the data and models under consideration.

Our simulation analysis has its limitations. The simulation parameters were generated using utilities derived from the EQ5D measured in a diabetic population in Canada. The distribution of utilities, especially the extent of the ceiling effect, will vary when studying different populations, or using a different measure of utility (e.g., the HUI). Populations who are sicker will exhibit a smaller ceiling effect, and thus the differences between the OLS, Tobit and CLAD models will likely diminish. Conversely, populations in better health will exhibit larger ceiling effects, which will increase the differences between the estimators. However, the philosophical point remains that it does not make sense to treat utilities as censored at 1 when the intent is to inform an economic evaluation. Similarly, whatever the population, the mathematical point remains that treating utilities as censored at 1 results in estimating the wrong parameters.

In this article, we have adopted the widely-held convention that the anchor points for health utilities are death and full health. We do, however, acknowledge that while death is a concrete state, each individual has their own notion of what constitutes full health. We also acknowledge that there is a school of thought whereby there exists a better-than-full health state which, it is argued, should have a utility that exceeds 1 (and hence, that an individual in this health state should accrue more than one QALY in a single year). A detailed discussion of this issue is beyond the scope of this article; however, we make the following observations. First, if better-than-full health states do indeed have utilities that exceed one, then when health utility is captured using the EQ5D or the HUI, the observed measurements are indeed censored. Second, if the outcome of interest is mean health utility (or mean QALYs), then neither the Tobit nor the CLAD model provides an acceptable solution. The CLAD models medians rather than means, and these two statistics will in general be different. The Tobit model assumes normality of the uncensored measurements, and it is impossible to demonstrate normality in the upper, unobserved tail of the utility distribution—we cannot assess normality in something we do not see. This motivates our third observation, namely that if such better-than-full health states exist, their associated utilities should be estimated through population TTO or SG studies in the same manner that utilities for other health states have been estimated. We reiterate, however, that the commonly-held interpretation is that utilities are bounded above at one, and that small departures from full health (and hence utilities which are just below one) are not captured by our measurement instruments. This results in a nonnegligible proportion of the population attaining the upper bound.

In conclusion, we note that Tobit and CLAD models, which assume that utilities can exceed 1, are not appropriate when the outcome of interest is a health utility that will be used in order to calculate QALYs for use in an economic analysis. OLS coupled with robust standard errors or the nonparametric bootstrap is asymptotically unbiased and produces valid confidence intervals. Moreover, it is easy to implement using standard software.

Source of financial support: EMP was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada. JET and DOR are career scientists supported by the Ministry of Health and Long-Term Care (Toronto, ON, Cnada). The original study on which the data in this study is based was funded by the Drug Innovation Fund administered by the Ministry of Health and Long-Term Care (Toronto, ON, Canada). The opinions expressed in the manuscript are our own and should not be attributed to any funding agency.

References

1 National Institute for Clinical Excellence (NICE). Guide to the Methods of Technology Appraisal. London, UK: NICE, 2004.
2 Guidelines for Economic Evaluation of Pharamaceuticals: Canada (2nd ed.). Ottawa: The Canadian Coordinating Office for Health Technology Assessment, 1997.
3 Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care 2002;40:113–28.
4 Torrance GW, Feeny DH, Furlong WJ, et al. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. Med Care 1996;34:702–22.
5 The EuroQol Group. EuroQol—a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.
6 Austin PC. A comparison of methods for analyzing health-related quality-of-life measures. Value Health 2002;5:329–37.
7 Austin PC, Escobar M, Kopec JA. The use of the Tobit model for analyzing measures of health status. Qual Life Res 2000;9:901–10.
8 Huang IC, Frangakis C, Atkinson MJ, et al. Addressing ceiling effects in health status measures: a comparison of techniques applied to measures for people with HIV disease. Health Serv Res 2008;43(1 Pt 1):327–39.
9 Gold MR, Siegel JE, Russell LB, Weinstein MC. Cost-Effectiveness in Health and Medicine. Oxford: Oxford University Press, 1996.
10 Ware JE, Kosinski M, Bjorner JB, Turner-Bowker DM, Gandek B, Maruish ME. Users' Manual for the SF-36v2 Health Survey (2nd ed.). Lincoln: Quality Metric, 2007.
11 Bellamy N, Buchanan WW, Goldsmith CH, et al. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988;15:1833–40.
12 Grootendorst P. Censoring in statistical models of health status: What happens when one can do better than "1". Qual Life Res 2000;9:911–14.
13 Thompson SG, Barber JA. How should cost data in pragmatic randomised trials be analysed? BMJ 2000;320:1197–200.
14 Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 2000;19:1141–64.
15 Walters SJ, Campbell MJ. The use of bootstrap methods for analysing health-related quality of life outcomes (particularly the SF-36). Health Qual Life Outcomes 2004;2:70.
16 Long JS, Ervin LH. Using heteroscedasticity consistent standard errors in the linear regression model. Am Stat 2000;54:217–24.
17 Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Series B Stat Methodol 1977;39:1–38.
18 Sullivan PW, Ghushchyan V. Preference-based EQ5D index scores for chronic conditions in the United States. Med Decis Making 2006;26:410–20.